Proteins and protein databases

Anne Bresciani

21/9/11

Background- Nucleotide databases

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/

National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), USA.

EMBL, http://www.ebi.ac.uk/embl/

European Bioinformatics Institute (EBI), England

(Established in 1980 by the European Molecular Biology Laboratory, Heidelberg, Tyskland)

DDBJ, http://www.ddbj.nig.ac.jp/

National Institue of Genetics, Japan

Together they form

International Nucleotide Sequence Database Collaboration, http://www.insdc.org/

Protein databases

Swiss-Prot, http://www.expasy.org/sprot/
Established in 1986 in Swizerland
ExPASy (Expert Protein Analysis System)
Swiss Institute of Bioinformatics (SIB) and European Bioinformatics Institute (EBI)

PIR, http://pir.georgetown.edu/
Established in 1984
National Biomedical Research Foundation, Georgetown University, USA

In 2002 merged to:
UniProt, http://www.uniprot.org/
A collaboration between SIB, EBI and Georgetown University.

UniProt

UniProt Knowledgebase (UniProtKB)
UniProt Reference Clusters (UniRef)
UniProt Archive (UniParc)

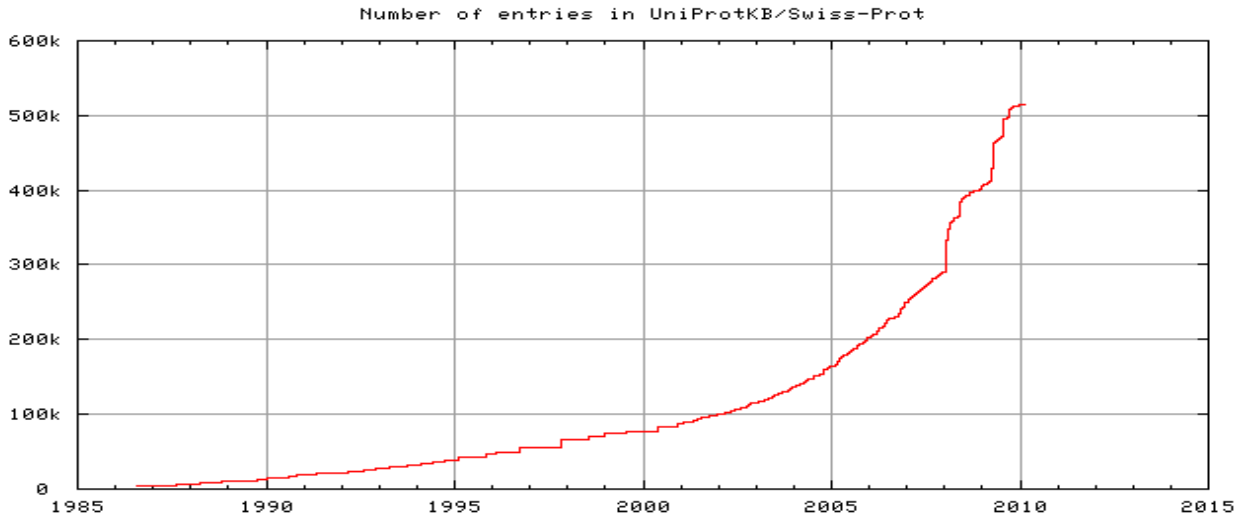
UniProt Knowledgebase Release 15.14 consists of:

UniProtKB/Swiss-Prot: Annotated manually (*curated*)
Release 57.14 of 09-Feb-2010: 514789 entries

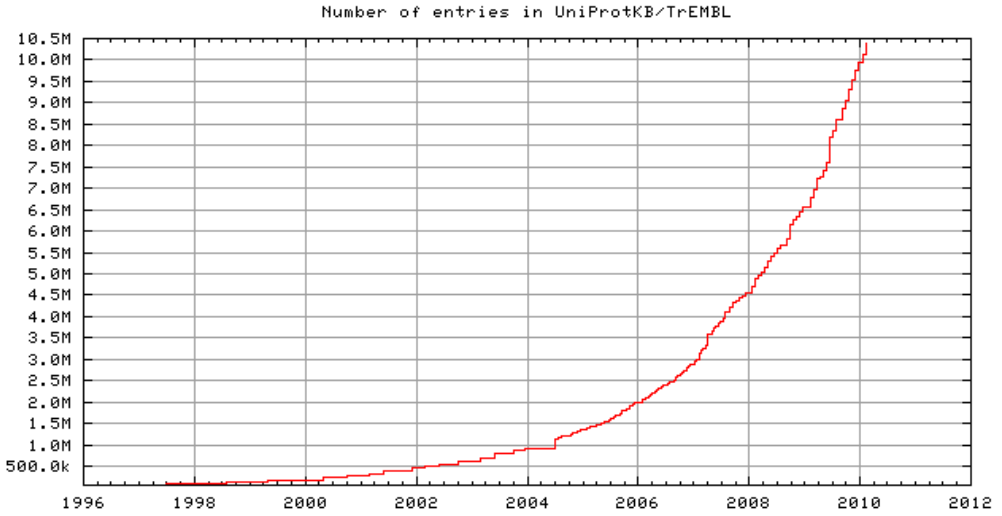
UniProtKB/TrEMBL: Computer annotated
Release 40.14 of 09-Feb-2010: 10376872 entries

Growth of UniProt

Swiss-Prot



TrEMBL



Content of UniProt Knowledgebase

- Amino acid sequences
- Functional and structural annotations
  - Function / activity
  - Secondary structure
  - Subcellular location
  - Mutations, phenotypes
  - Post-translational modifications
- Origin
  - organism: Species, subspecies; classification
  - tissue
- References
- Cross references

Amino acid sequences

From where do you get amino acid sequences?

- Translation of nucleotide sequences (GenBank/ EMBL/DDBJ)
- Amino acids sequencing: *Edman degradation*
- Mass spectrometry
- 3D-structures

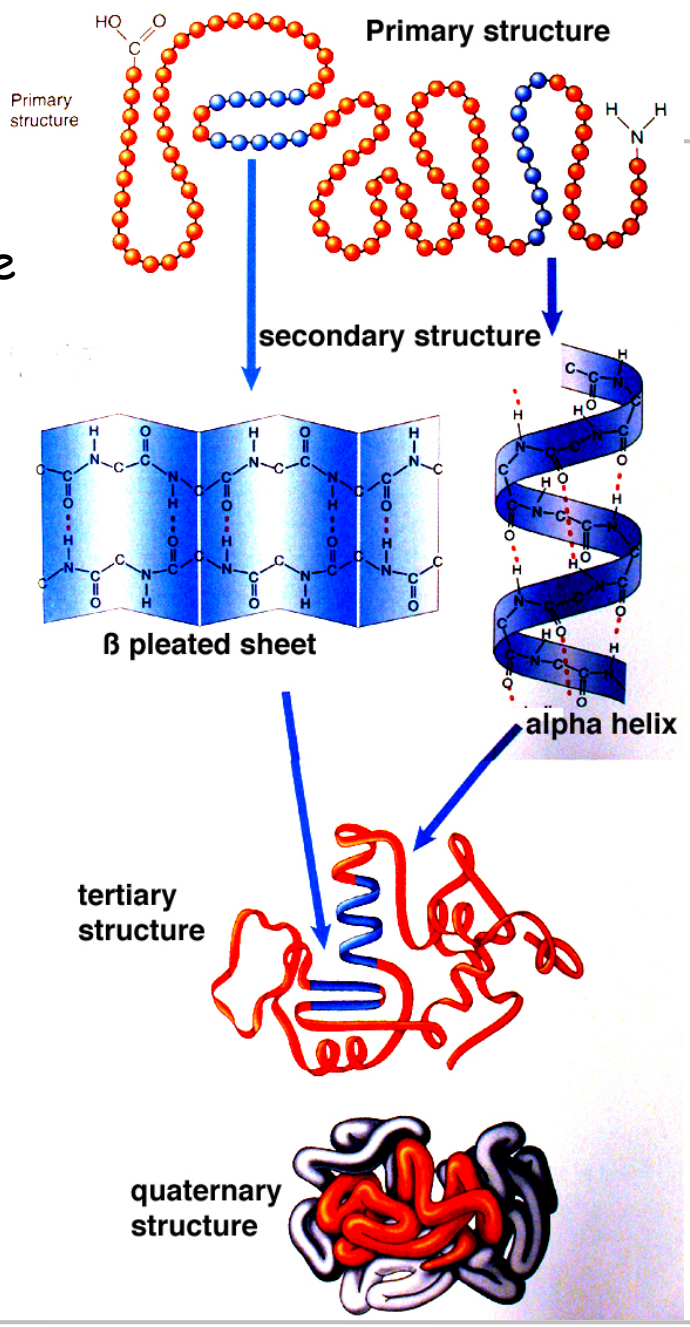
Protein structure

Primary structure: Amino acid sequence

Secondary structure:
"Backbone" hydrogen bonding
Alpha helix / Beta sheet / Turn

Tertiary structure: Fold, 3D coordinates

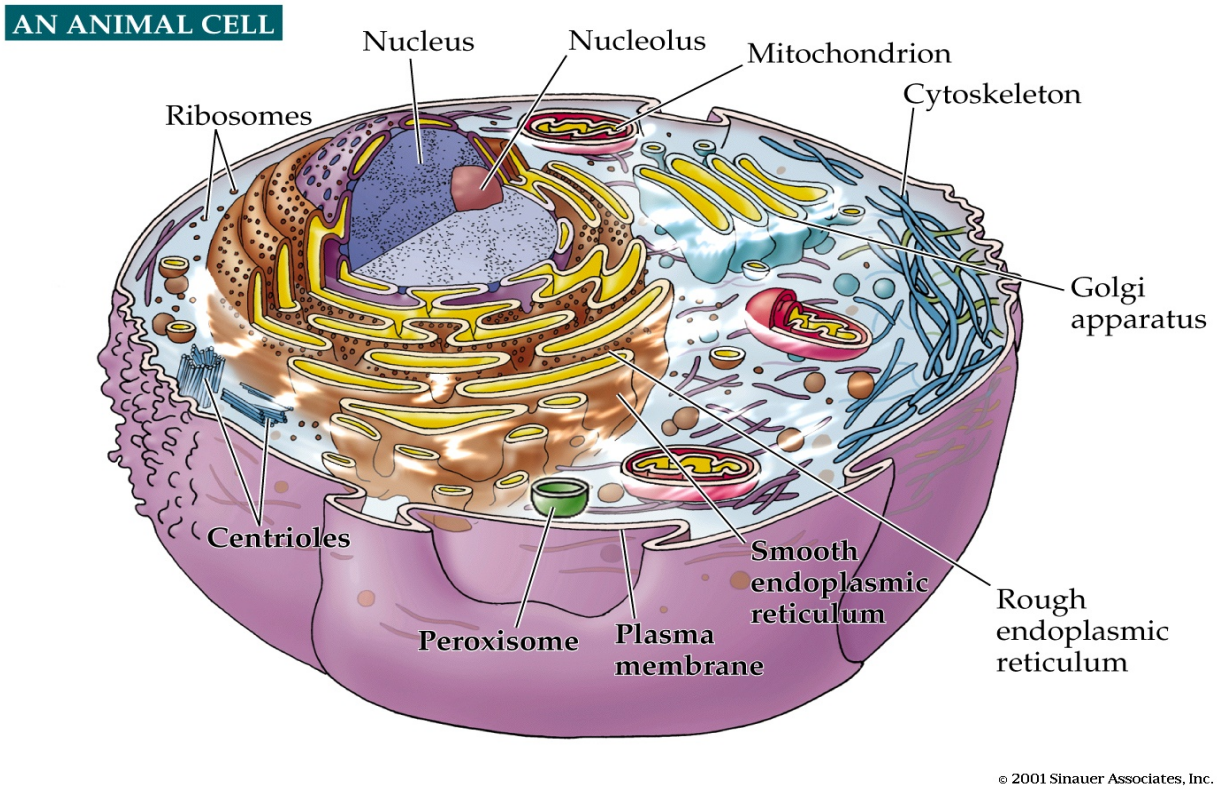
Quaternary structure: subunits



Subcellular location

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

An animal cell:



Post-translational modifications

- Cleavage of signal peptide, transit peptide or pro-peptide
- Phosphorylation
- Glycosylation
- Lipid anchors
- Disulfide bond
- Prosthetic groups (*e.g.* metal ions)

Cross references

Other databases (there are 95 in total):

- Nucleotide sequences
- 3D structure
- Protein-protein interactions
- Enzymatic activities and pathways
- Gen expression (microarrays and 2D-PAGE)
- Ontologies
- Families and domains
- Organism specific databases

Evidence

3 types of *non-experimental qualifiers* in Sequence annotation and General comment:

- *Potential:* Predicted using sequence analysis
- *Probable:* Uncertain experimental evidence
- *By similarity:* Predicted using sequence similarity

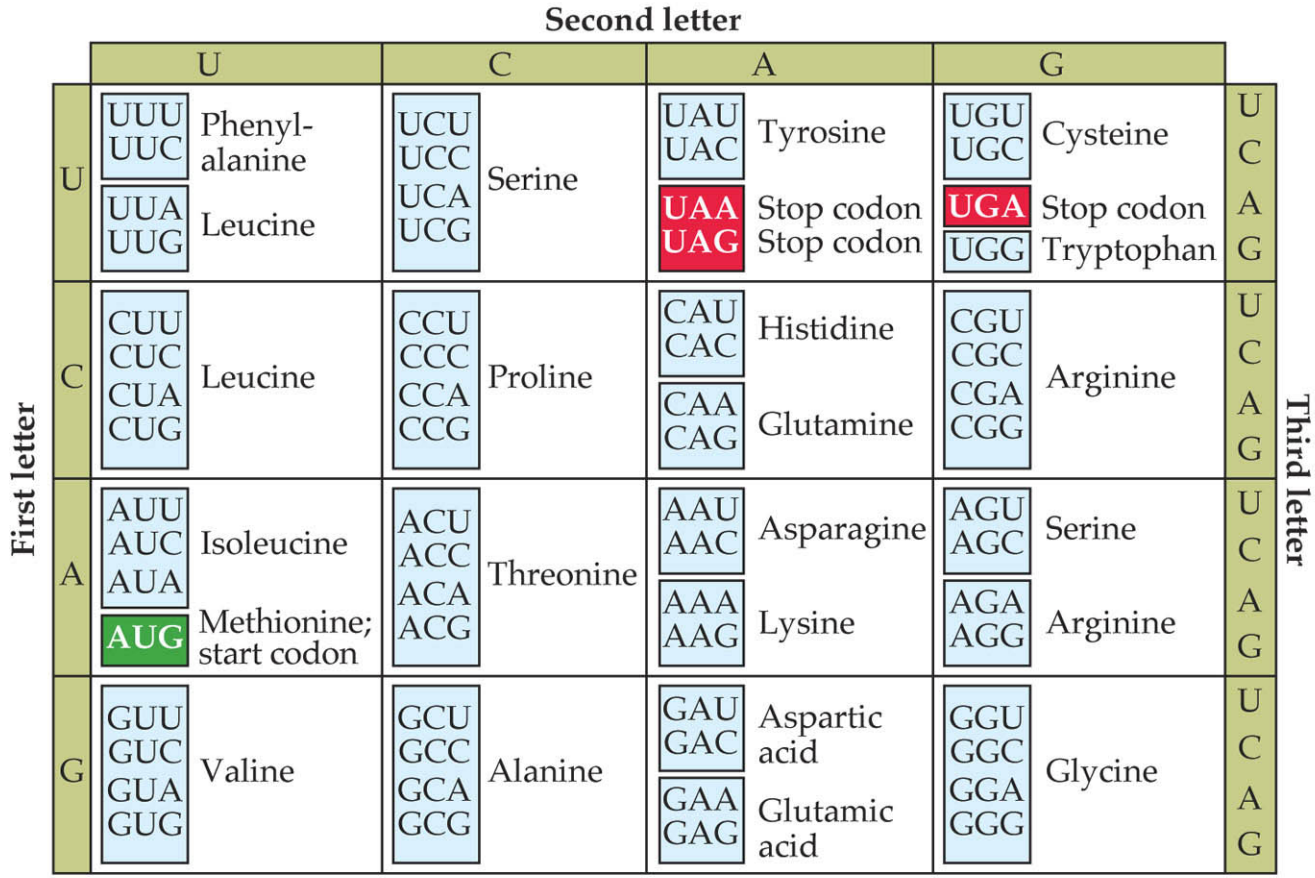
Content of UniProt Knowledgebase

http://www.uniprot.org/uniprot/P00299

Translation

- The central dogma
- How do we go from RNA to protein

The genetic code



- Degenereret (*redundant*) men ikke tvetydig (*ambiguous*)
- *Næsten* universel (afvigelser er fundet i mitokondrier)

Læserammer 1

Et stykke af en mRNA-streng:

```
5'                                                              3'
augcccaagcugaauagcguagaggggguuuucaucauuugaggacgauguauaa
```

kan opdeles i tripletter (*codons*) på tre måder:

```
1 aug ccc aag cug aau agc gua gag ggg uuu uca uca uuu gag gac gau gua uaa
   M   P   K   L   N   S   V   E   G   F   S   S   F   E   D   D   V   *
2  ugc cca agc uga aua gcg uag agg ggu uuu cau cau uug agg acg aug uau
    C   P   S   *   I   A   *   R   G   F   H   H   L   R   T   M   Y
3   gcc caa gcu gaa uag cgu aga ggg guu uuc auc auu uga gga cga ugu aua
     A   Q   A   E   *   R   R   G   V   F   I   I   *   G   R   C   I
```

Hver mulig opdeling kaldes en *læseramme (reading frame).*

Læserammer 2

Eftersom der er to strenge i DNA, er der *seks* mulige læserammer på et stykke DNA (tre I hver retning):

```
 3    A  Q  A  E  *  R  R  G  V  F  I  I  *  G  R  C  I
 2   C  P  S  *  I  A  *  R  G  F  H  H  L  R  T  M  Y
 1  M  P  K  L  N  S  V  E  G  F  S  S  F  E  D  D  V  *
5' ATGCCCAAGCTGAATAGCGTAGAGGGGTTTTCATCATTTGAGGACGATGTATAA 3'
3' TACGGGTTCGACTTATCGCATCTCCCCAAAAGTAGTAAACTCCTGCTACATATT 5'
    H  G  L  Q  I  A  Y  L  P  K  *  *  K  L  V  I  Y  L  -1
      G  L  S  F  L  T  S  P  N  E  D  N  S  S  S  T  Y   -2
     A  W  A  S  Y  R  L  P  T  K  M  M  Q  P  R  H  I    -3
```

En læseramme fra et startcodon til det første stopcodon kaldes en *åben* læseramme (understreget ovenfor).